An Empirical Study of Catastrophic Forgetting in Large Language Models During Continual Fine-Tuning

Year: 2023 | Citations: 476 | Authors: Yun Luo, Zhen Yang, Fandong Meng

---

Abstract

Catastrophic forgetting (CF) is a phenomenon that occurs in machine learning when a model forgets previously learned information while acquiring new knowledge for achieving satisfactory performance in downstream tasks. As large language models (LLMs) have demonstrated remarkable performance, it is intriguing to investigate whether CF exists during the continual instruction tuning of LLMs. This study empirically evaluates the forgetting phenomenon in LLMs' knowledge during continual instruction tuning from the perspectives of domain knowledge, reasoning, and reading comprehension. The experiments reveal that catastrophic forgetting is generally observed in LLMs ranging from 1 b to 7 b parameters. Surprisingly, as the model scale increases, the severity of forgetting intensifies in such a model scale range, which may result from the much more significant initial performance in the larger LLM. The finding is also observed by the experiment of Qwen-2.5-Inst from 3 B to 14 B. Comparing the decoder-only model BLOOMZ with the encoder-decoder model mT0, BLOOMZ exhibits less forgetting and retains more knowledge. Interestingly, we also observe that LLMs can mitigate language biases, such as gender bias, during continual fine-tuning. Furthermore, our findings indicate that general instruction tuning can help alleviate the forgetting phenomenon in LLMs during subsequent fine-tuning.